

Hybridization Number on Three Trees

Leo van Iersel ^{*†} Steven Kelk [‡] Nela Lekić ^{†‡} Chris Whidden [§] Norbert Zeh [¶]

May 2, 2014

Abstract

Phylogenetic networks are leaf-labelled directed acyclic graphs that are used to describe non-treelike evolutionary histories and are thus a generalization of phylogenetic trees. The hybridization number of a phylogenetic network is the sum of all indegrees minus the number of nodes plus one. The Hybridization Number problem takes as input a collection of phylogenetic trees and asks to construct a phylogenetic network that contains an embedding of each of the input trees and has a smallest possible hybridization number. We present an algorithm for the Hybridization Number problem on three binary phylogenetic trees on n leaves, which runs in time $O(c^k \text{poly}(n))$, with k the hybridization number of an optimal network and c some positive constant. For the case of two trees, an algorithm with running time $O(3.18^k n)$ was proposed before whereas an algorithm with running time $O(c^k \text{poly}(n))$ for more than two trees had prior to this article remained elusive. The algorithm for two trees uses the close connection to acyclic agreement forests to achieve a linear exponent in the running time, while previous algorithms for more than two trees (explicitly or implicitly) relied on a brute force search through all possible underlying network topologies, leading to running times that are not $O(c^k \text{poly}(n))$, for any c . The connection to acyclic agreement forests is much weaker for more than two trees, so even given the right agreement forest, the reconstruction of the network poses major challenges. We prove novel structural results that allow us to reconstruct a network without having to guess the underlying topology. Our techniques generalize to more than three input trees with the exception of one key lemma that maps nodes in the network to tree nodes and, thus, minimizes the amount of guessing involved in constructing the network. The main open problem therefore is to prove results that establish such a mapping for more than three trees.

^{*}Centrum Wiskunde & Informatica (CWI), Science Park 123, 1098 XG Amsterdam, The Netherlands, l.j.v.iersel@gmail.com

[†]Leo van Iersel and Nela Lekić were respectively funded by a Veni and a Vrije Competitie grant from The Netherlands Organisation for Scientific Research (NWO).

[‡]Department of Knowledge Engineering (DKE), Maastricht University, PO. Box 616, 6200 MD Maastricht, The Netherlands, steven.kelk@maastrichtuniversity.nl, nela.lekic@maastrichtuniversity.nl

[§]Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., PO Box 19024, Seattle, WA, USA 98109, whidden@cs.dal.ca

[¶]Faculty of Computer Science, Dalhousie University, 6050 University Ave, Halifax, NS B3H 1W5, Canada, nzeh@cs.dal.ca

1 Introduction

In computational biology the evolutionary history of a set of contemporary species (or *taxa*) is often modelled as a *rooted phylogenetic tree*. Informally this is a rooted tree in which the leaves are bijectively labelled by the taxa and edges are directed away from the root, reflecting the direction of evolution [18]. Nodes of outdegree two or higher model the points in history at which a common ancestor of a subset of the taxa differentiated into two or more sublineages. The central problem in phylogenetics is to recover the topology of the “true” phylogenetic tree, given only information about the taxa, often DNA data. This is a challenging computational problem and has been the topic of intensive research during the last 40 years [8]. Recently our understanding of evolutionary mechanisms has deepened and there is growing awareness that evolution is not always treelike [1]. In particular, due to *reticulate phenomena* such as hybridization and horizontal gene transfer [15] the evolution of a set of species is sometimes better modelled as a *rooted phylogenetic network* [9], essentially a generalisation of phylogenetic trees to directed acyclic graphs (DAGs). In such graphs nodes with indegree two or higher, known as *reticulations*, represent the points at which two or more lineages merge, rather than diversify.

The study of rooted phylogenetic networks is comparatively new and it has given rise to many novel and hard combinatorial optimization problems [9]. In this article we focus on the *hybridization number* problem, originally introduced in [2, 3], which is one of the most well-studied phylogenetic network problems to date. Here we are given a set of rooted phylogenetic trees \mathcal{T} , on the same set of taxa X , and the goal is to construct a phylogenetic network - henceforth called a *hybridization network* - that contains an image of each of the input trees, while minimizing the hybridization number k of the network. If we restrict (without loss of generality) to networks with maximum indegree two, the hybridization number is simply equal to the number of reticulation nodes. We defer exact definitions to the preliminaries. See Figure 1 for an example of a hybridization network (with hybridization number three) for three input trees.

The holy grail for this problem is to develop algorithms that can cope with many input trees and nonbinary input trees [15] (and to take different causes of incongruence into account, see e.g. [24]). However, thus far most algorithmic research has focussed on the simplest possible case: $|\mathcal{T}| = 2$ and both input trees are binary. Unfortunately even this version of the problem is NP-hard and APX-hard [5], with similar (in)approximability properties to the classical problem Directed Feedback Vertex Set [13]. Fortunately the binary two-tree problem is Fixed Parameter Tractable (FPT) in k . (See [7, 16] for an introduction to fixed parameter tractability). This result was initially established via kernelization - the problem has a quadratic kernel [4] - but the theoretical state of the art is an algorithm based on bounded-search with running time $O(3.18^k \cdot \text{poly}(n))$ [22] with $n = |X|$. The comparative tractability of the problem, both in theory and practice (see e.g. [6] for a fast implementation), stems from the essentially one-to-one relationship between solutions to the two-tree problem and the Maximum Acyclic Agreement Forest (MAAF) problem. In the latter problem (originally introduced in [3]) one is required to cut the two input trees into common components, such that the number of components is minimized and there are no cyclical dependencies between components. The MAAF abstraction gives a useful static characterization of the two-tree hybridization number problem [5]. In particular, in the two-tree case the

MAAF abstraction essentially allows us to bypass the problem of actually constructing the hybridization network: it can easily be constructed in polynomial time from the components of the MAAF. The MAAF abstraction, and related FPT results, also hold in the case of two *nonbinary* trees, albeit with significant technical complications [14, 17].

For $|\mathcal{T}| > 2$ the situation becomes more complex, however, even when restricted to binary trees¹ and $|\mathcal{T}| = 3$. The MAAF abstraction weakens significantly and cannot (obviously) be used to generate optimal solutions to the hybridization number problem. Without the MAAF abstraction it seems that we have to explicitly confront the challenge of actually constructing the hybridization network itself. This is a theoretically daunting challenge, since the space of DAGs is huge. The good news is that for $|\mathcal{T}| > 2$ the problem nevertheless remains FPT in k [11, 21]. The bad news is that none of these results satisfactorily address the problem of actually constructing the network. The FPT result in [21] gives a quadratic kernel, but does not describe a (good) algorithm for solving the kernel. The bounded-search FPT result in [11], based on [12], does actually construct the network, but has an astronomical running time. The running time is so large because it brute forces over the space of all possible *generators* i.e. possible “backbone topologies” of the network [10, 12], a space which is not known to be $O(c^k)$, and continues with a tower of guesses, which is not $O(c^k)$, for each such generator. At present, therefore, the only FPT algorithms for the case of three binary trees are either kernelizations, or bounded-search algorithms with an exponential dependency on k with a non-linear exponent. Several exponential-time algorithms do exist, such as [23] and the algorithm discussed in [21], but using them to solve a kernelized hybridization number instance unfortunately does not help for two reasons. Firstly, the size of the best-known kernel (i.e. the number n of leaves of a kernelized instance) is quadratic, and not linear in k . Secondly, no previously-known exponential-time algorithm has an $O(c^n)$ running time. Therefore, the challenge is to determine whether an algorithm with running time $O(c^k \cdot \text{poly}(n))$ exists for the case of three binary trees.

In this article we answer this challenge positively. Although the constant c that we find is astronomical - 1609891840 - it represents a significant development in our understanding of the underlying combinatorial structure of the hybridization number problem. We show that, although it is not clear how a MAAF can be pieced together into an optimal solution to the hybridization number problem, it is still possible to identify in $O(c^k \text{poly}(n))$ time a (not necessarily maximum) acyclic agreement forest that does have this property. Having found the appropriate acyclic agreement forest, we use deep insights into the structure of optimal hybridization networks to piece the components of the forest together into a network. The difficulty of this step comes from the fact that, unlike in the two-tree case, it is no longer possible to avoid having nodes in the network that are separated from all leaves by hybridization edges, and that are hence not represented in the agreement forest. The main insight helping to overcome this problem is that, in the case $|\mathcal{T}| = 3$, there always exists an optimal hybridization network such that each of its outdegree-2 nodes corresponds to one or more outdegree-2 nodes of the input trees, see Figure 1. This enables us to keep the combinatorial explosion in the number of possible network topologies under

¹For the rest of the introduction we focus only on the case of binary trees – see [19] and [20] for an overview of recent nonbinary results. The nonbinary case is a generalization of the binary case and therefore inherits all the negative results, but not the positive results, of the binary problem.

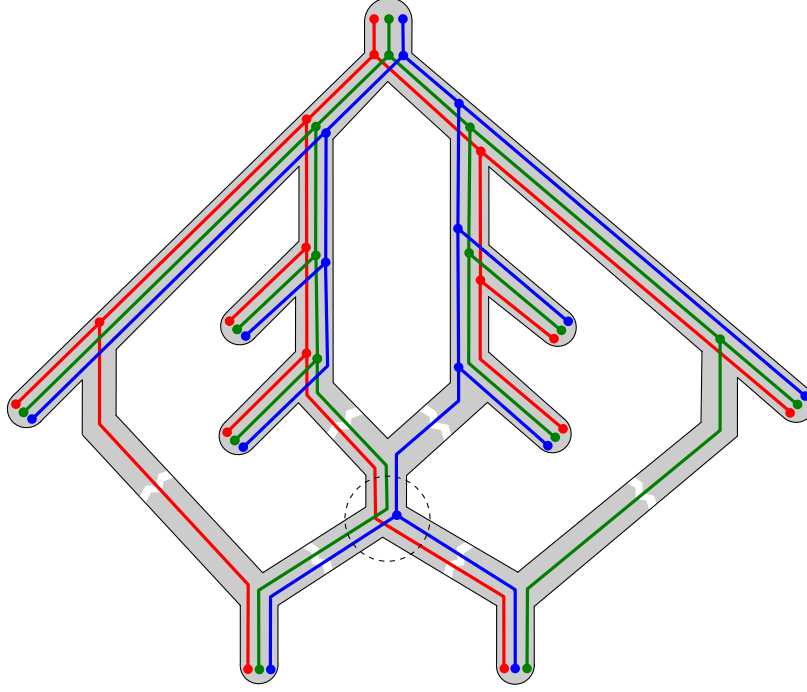


Figure 1: A hybridization network for three trees which contains an invisible component (inside the dashed circle). It can be shown that any hybridization network for these trees contains an invisible component. However, the single node inside this component can be identified because it corresponds to a node of the blue input tree.

control.

Note that our algorithm can be viewed as a structural generalisation of existing algorithms for two trees, which also separate the identification of the underlying acyclic agreement forest and the construction of the network into two phases. In the case of two trees the second phase is polynomial and it is comparatively easy to obtain $O(c^k \text{poly}(n))$ running times for the first phase. In fact, although our overall result at present only holds for the case $|\mathcal{T}| = 3$, the results for the first phase go through without modification for the case $|\mathcal{T}| > 3$. As we demonstrate, the only barrier to extending our result is the fact that, for $|\mathcal{T}| > 3$, the combinatorial insight mentioned in the previous paragraph no longer holds. Indeed, there are two new challenges stemming from this article. Firstly, to adapt and generalize the combinatorial insight so that the wider result can be extended to four or more trees. Secondly, to significantly optimize the constant c in our running time. How close can we get to the competitive $O(3.18^k \cdot \text{poly}(n))$ running time achieved in the two tree case?

The structure of the remainder of the article is as follows. In Section 2, we present the necessary definitions. Section 3 shows how we can “guess” the underlying acyclic agreement forest of an optimal hybridization network in $O(c^k \text{poly}(n))$ time. Then we define a notion of canonical networks in Section 4 (basically, networks of which each outdegree-2 node corresponds to an outdegree-2 node of at least one of the input trees) and show that we may restrict to canonical networks as long as there are at most three trees in the input. Subsequently, Section 5 then shows how such a canonical network can be

reconstructed from an acyclic agreement forest and the input trees in $O(c^k \text{poly}(n))$ time. Finally, we present our conclusions in Section 6 and give an example of the algorithm in the appendix.

2 Preliminaries

A *rooted phylogenetic X -tree* is a rooted tree with no nodes with indegree 1 and outdegree 1, a root with indegree 0 and outdegree 1, and leaves bijectively labeled by the elements of a finite set X . Such a tree is called *binary* if all inner nodes except the root have indegree 1 and outdegree 2. From now on we will refer to a rooted binary phylogenetic X -tree as a *tree* for short.

A *rooted phylogenetic network* is a directed acyclic graph (DAG) with no nodes with indegree 1 and outdegree 1, a single indegree 0 node (the *root*) with outdegree 1 and leaves bijectively labelled by the elements of X . Rooted phylogenetic networks will be called *networks* for short. We identify each leaf of a tree or network with its label and call directed edges *edges* for short. We see the root of a tree or network as a leaf and assume without loss of generality that it is labelled ρ . We call a network *binary* if every non-leaf node has total degree 3 and all leaves have degree 1.

Network nodes with indegree 1 and outdegree at least 2 we call *split nodes* while nodes with indegree at least 2 we call *reticulation nodes*, or *reticulations* for short. The *hybridization number* (often also called *reticulation number*) of a binary network N is defined as the number of reticulation nodes of N . For a general network the hybridization number is given by the sum $\sum (d^-(v) - 1)$ over all nodes v of N with indegree $d^-(v)$ at least 2. For a tree T and a set $X' \subseteq X$, we define $T(X')$ as the minimal subtree of T that contains all elements of X' , and as the result of suppressing all nodes of $T(X')$ with in- and outdegree 1. The set of leaves of a tree T is denoted $L(T)$.

We say that a tree T is *displayed* by a network N if T can be obtained from a subgraph of N by contracting edges. Given a set \mathcal{T} of rooted phylogenetic trees, the MINIMUM HYBRIDIZATION problem asks to find a phylogenetic network N that displays each tree in \mathcal{T} such that the hybridization number of N is minimized. We say that N is a *hybridization network* for a set \mathcal{T} of input trees if N displays all $T \in \mathcal{T}$. In addition, we say that the *hybridization number* of a set of input trees \mathcal{T} is the hybridization number of a hybridization network for \mathcal{T} that has the lowest hybridization number over all hybridization networks for \mathcal{T} . It is well known that if there exists a hybridization network for \mathcal{T} , there also exists a binary hybridization network for \mathcal{T} with the same hybridization number (see e.g. [11]). Therefore, all hybridization networks are from now on assumed to be binary.

Let T be a rooted phylogenetic X -tree and S a rooted phylogenetic X' -tree for some $X' \subseteq X$. We say that S is a *pendant subtree* of T if it is a subtree that can be detached from T by deleting a single edge. For a set \mathcal{T} of phylogenetic X -trees and $X' \subseteq X$, a *common pendant subtree* of \mathcal{T} is a rooted phylogenetic X' -tree that is a pendant subtree of each tree in \mathcal{T} . Let T be a tree, let (x_1, x_2, \dots, x_q) be a tuple of elements of X with $q \geq 1$ and let p_i denote the parent of x_i in T . We say that the tuple (x_1, x_2, \dots, x_q) is a *chain* of T if either $(p_q, p_{q-1}, \dots, p_1)$ is a directed path in T , or $(p_q, p_{q-1}, \dots, p_2)$ is a directed path in T and $p_1 = p_2$. A *common chain* of a set \mathcal{T} of trees is a maximal tuple (x_1, x_2, \dots, x_q) that is a chain of each tree in \mathcal{T} .

Related to the hybridization number problem is a concept of agreement forests. A forest is a collection of trees, which we will call components rather than trees to avoid confusion with the input trees. We say that a forest \mathcal{F} is a *forest for* a tree T if $T|L(F)$ is isomorphic to F for all $F \in \mathcal{F}$ and the trees $\{T(L(F)) \mid F \in \mathcal{F}\}$ are node-disjoint subtrees of T whose leaf-set union equals $L(T)$. By this definition, if \mathcal{F} is a forest for some tree T , then $\{L(F) \mid F \in \mathcal{F}\}$ is a partition of the leaf set of T . It will indeed sometimes be useful to see a forest as a partition of the leaves and sometimes to see it as a collection of trees. If \mathcal{T} is a set of trees, then a forest \mathcal{F} is an *agreement forest* of \mathcal{T} if it is a forest for each $T \in \mathcal{T}$. Note that these definitions only apply to binary trees.

We define the *inheritance graph* $IG(\mathcal{T}, \mathcal{F})$ of an agreement forest \mathcal{F} of a set \mathcal{T} of trees as the directed graph whose node set is the set of components of \mathcal{F} and whose edge set contains an edge (F, F') precisely if there is a directed path from the root of $T(L(F))$ to the root of $T(L(F'))$ in at least one $T \in \mathcal{T}$. An agreement forest \mathcal{F} of \mathcal{T} is called an *acyclic agreement forest* (AAF) if the graph $IG(\mathcal{T}, \mathcal{F})$ does not contain any directed cycles.

The last definition we need is the notion of generators (see e.g. [12]), which we use to describe the underlying structure of networks without nontrivial pendant subtrees. A (binary) *r-reticulation generator* is defined as an acyclic directed multigraph with a single root with indegree 0 and outdegree 1, exactly r nodes with indegree 2 and outdegree at most 1, and all other nodes have indegree 1 and outdegree 2. If N is a network, then the *underlying generator* of N is the generator obtained from N by deleting all leaves and suppressing all indegree 1 outdegree 1 nodes. The *sides* of a generator are its edges (the *edge sides*) and its nodes with indegree 2 and outdegree 0 (the *node sides*). Thus, each leaf of a network N is on a certain side of its underlying generator. See Figure 2 for an example.

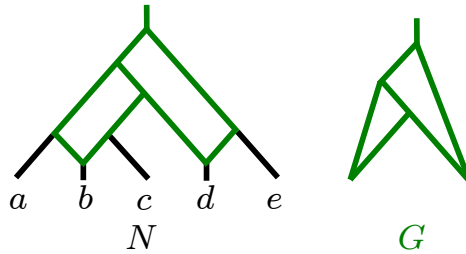


Figure 2: Graph G is the generator underlying network N . It has 9 sides: 7 edge sides and 2 node sides.

3 Guessing the AAF

Let \mathcal{T} be a collection of input trees. Without loss of generality we will assume that \mathcal{T} contains no nontrivial common pendant subtrees. In this section, we show how we can guess an AAF from which we can build an optimal hybridization network for \mathcal{T} . To make this precise, we define the *deletion forest* of a network N as the forest obtained by deleting all the reticulation edges of N , deleting all resulting connected components that do not contain any taxa, and then taking the partition of the taxa induced by the remaining connected components. Note that for a given network the deletion forest is uniquely defined. We start by proving the following lemma.

Lemma 1. *Given a hybridization network N with hybridization number k for a set \mathcal{T} of input trees, the deletion forest of N is an AAF of \mathcal{T} with at most $k + 1$ components.*

Proof. We first show that the deletion forest contains at most $k + 1$ components. To see this, note that N contains exactly k reticulation nodes. For a reticulation node r , let $X(r)$ be the set of taxa that can be reached from r by directed paths that start at r and which do not intersect with any reticulation apart from r . (Possibly, $X(r) = \emptyset$.) By construction, none of the edges on these directed paths are deleted when the deletion forest is created. Hence, all the taxa in $X(r)$ will be in the same connected component. Similarly, if $X(\rho)$ denotes the set of taxa reachable by directed paths that start at the root and which do not intersect with any reticulations, then the taxa in $X(\rho)$ will also be together in the same connected component. Note that the deletion forest \mathcal{F} of N (seeing it as a partition of the taxa) is the collection containing $X(\rho)$ if $X(\rho) \neq \emptyset$ and $X(r)$ for each reticulation r for which $X(r) \neq \emptyset$. Hence, the deletion forest contains at most $k + 1$ components. Moreover, for each $F \in \mathcal{F}$, each input tree must yield the same subtree when restricted to the subset of taxa of F because N displays all the input trees. In addition, for each input tree $T \in \mathcal{T}$, the subtrees $\{T(L(F)) \mid F \in \mathcal{F}\}$ are node-disjoint, again because N displays T . It follows that the deletion forest \mathcal{F} of N is indeed an AAF of the input trees, with at most $k + 1$ components. \square

As a consequence of Lemma 1, we will from now on refer to the deletion forest of a network as its *deletion AAF*. We can now show how one can guess the deletion AAF of some optimal hybridization network for the input trees.

Lemma 2. *Let k be the hybridization number of the set \mathcal{T} of input trees. Then, in time $O(c^k \cdot \text{poly}(n))$, we can guess an AAF that is the deletion AAF of some hybridization network for \mathcal{T} with hybridization number k .*

Proof. Consider an arbitrary input tree $T \in \mathcal{T}$. Observe that an AAF of \mathcal{T} with $k' + 1$ components can be obtained from T by deleting exactly k' edges and taking the partition of the taxa induced by the resulting connected components. By Lemma 1, the deletion AAF of an optimal hybridization network for \mathcal{T} has at most $k + 1$ components. The goal therefore is to locate the at most k edges that need to be deleted from T in order to obtain the deletion AAF of some optimal hybridization network for \mathcal{T} .

Let N be a hybridization network for \mathcal{T} with hybridization number k and consider its underlying generator, which is a k -reticulation generator and hence has at most k node sides and at most $4k - 1$ edge sides [21]. It follows that there are at most $5k - 1$ common chains of \mathcal{T} , because any two taxa on the same edge side of N are in the same common chain [21]. The set of common chains is unambiguously defined by the set of input trees, can be computed in polynomial time, and no two chains can share a taxon. Moreover, in [21] it is proven that if two or more taxa of a common chain are on a single edge side of the underlying generator, the entire chain can safely be moved onto that edge side. That is, the new network still displays the input trees and has hybridization number no higher than the old network.

This means that we can assume the existence of an optimal network N' such that for each common chain there are exactly two possibilities: (1) the chain is on a single side of the underlying generator, or (2) each taxon of the chain is on a different side of the underlying generator. For each chain we can

guess which of the two cases holds, using at most 2^{5k-1} guesses for the entire set of chains. Since, as mentioned before, any two taxa that are on the same side are together in a common chain, it follows that each side of (the underlying generator of) N' contains either a complete case (1) chain or a single taxon (which is either in a case (2) chain or a singleton-chain) or no taxa at all.

Now, assume that we have identified the correct set of guesses describing the behaviour of the common chains in N' . It remains to show that we can identify the correct set of edges to delete in T to obtain the deletion AAF corresponding to N' . Observe that for each case (1) chain it is not necessary to delete any of the internal edges of the chain in T . This is because we have correctly identified that the entire chain sits on a single edge side of the generator and thus that it sits inside a single component of the deletion AAF. For all other edges in T we can simply guess whether to delete it or not. Fortunately, there are not too many of these edges. Specifically, recall that each side contains either a case (1) chain or a single taxon, and that the number of sides is at most $5k - 1$. Hence, if we collapse each case (1) chain C into a new taxon x_C , which is permitted because we will never cut its internal edges, there are in total at most $5k - 1$ taxa left. A binary tree with $5k - 1$ taxa has $10k - 4$ edges. By brute forcing on these edges (i.e. guessing whether to delete it or not) we observe that, in total, the deletion AAF of N' can be located in time at most $O(2^{5k} \cdot 2^{10k} \cdot \text{poly}(n))$. \square

4 Canonical networks

In this section we give the only lemma that is three-tree specific. We describe a transformation from a hybridization network to a structure, called a canonical network with embedded trees, that has some desirable properties. We prove that the transformation preserves the hybridization number and, in the three-tree case, we are thus allowed to concentrate on canonical networks alone. For ease of notation, we will sometimes identify a directed graph with its edge set.

A *Canonical Network with Embedded Trees (CNET)* for a set \mathcal{T} of phylogenetic trees over a label set X is a pair $\mathcal{H} = (H, \mathcal{E})$ with the following properties:

- (i) H is a DAG. We call its sources *roots* and its sinks *leaves*.
- (ii) Every root of H has one child.
- (iii) The leaves of H are labelled bijectively with the leaf labels in X .
- (iv) $\mathcal{E} = \{H(T) \subseteq H \mid T \in \mathcal{T}\}$ and, for all $T \in \mathcal{T}$, $H(T)$ is an *image* of T , that is, T can be obtained from $H(T)$ by suppressing degree-2 nodes.
- (v) Every tree image $H(T) \in \mathcal{E}$ contains an edge incident to a root of H .
- (vi) $H = \bigcup_{T \in \mathcal{T}} H(T)$, that is, every edge of H belongs to at least one tree image in \mathcal{E} .
- (vii) Every non-leaf non-root node of H has exactly two children.
- (viii) For every non-leaf non-root node, there exists a tree image $H(T) \in \mathcal{E}$ that contains both its child edges.

We represent the tree images in \mathcal{E} by associating a unique colour with each tree $T \in \mathcal{T}$ and colouring every edge in $H(T)$ with this colour. We call the colour associated with tree T *colour* T . We use $C(e)$ to

denote the colour set of an edge e of H , that is, the set of trees $T \in \mathcal{T}$ whose images $H(T) \in \mathcal{E}$ include e . A CNET for the input trees in Figure 10 on page 22 is shown in Figure 13 on page 23. A corresponding hybridization network is shown in Figure 14 on page 24.

The hybridization network *induced* by a CNET (H', \mathcal{E}) is the hybridization network obtained by applying the following transformations to H' :

- We replace every node x that is both a reticulation node and a split node with two nodes x_t and x_b , change the bottom endpoints of x 's parent edges to x_t , change the top endpoints of x 's child edges to x_b , and add an edge from x_t to x_b .
- We replace every reticulation node with more than two parents with a chain of binary reticulation nodes. See Figure 3 for an illustration of these first two steps.
- As long as there are at least two roots, we choose two such roots r_1 and r_2 , change the top endpoint of r_1 's child edge to r_2 , and add an edge from r_1 to r_2 . This reduces the number of roots by one, so we eventually obtain a network with a single root. See Figure 4.
- For every leaf x with more than one parent, we create a new node x' , change the bottom endpoint of every parent edge of x to x' , and add an edge from x' to x .

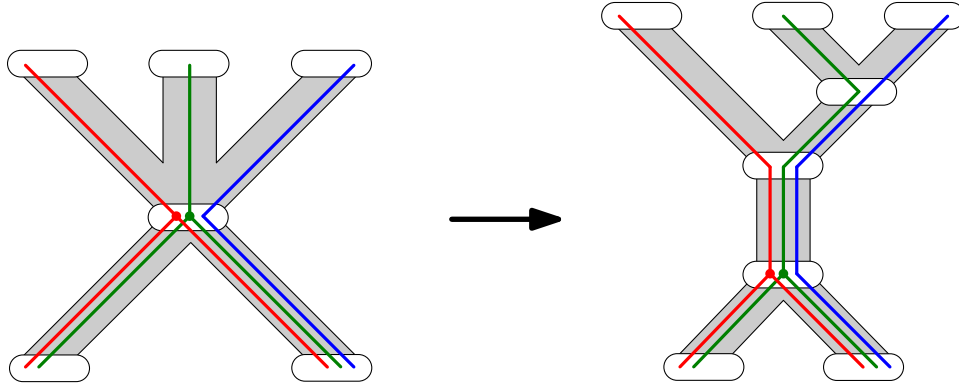


Figure 3: The first two steps of transforming a CNET into a hybridization network: expanding nodes that are both reticulation and split nodes and refining reticulations.

The *deletion AAF* of a CNET (H, \mathcal{E}) is defined to be the deletion AAF of the hybridization network induced by (H, \mathcal{E}) . The *hybridization number* of a CNET (H, \mathcal{E}) is (as for networks) defined to be the sum $\sum (d^-(v) - 1)$ over all nodes v of H with indegree $d^-(v)$ at least 2.

Lemma 3. *If $|\mathcal{T}| = 3$, then there exists a hybridization network H with hybridization number k for \mathcal{T} if and only if there exists a CNET $\mathcal{H} = (H', \mathcal{E})$ with hybridization number k for \mathcal{T} . Moreover, if H and \mathcal{H} exist then they have the same deletion AAF.*

Proof. First suppose that there exists a CNET $\mathcal{H} = (H', \mathcal{E})$ with hybridization number k for \mathcal{T} . Then, the hybridization network H induced by \mathcal{H} has the same hybridization number as H' , and it is easy to see that H displays the trees in \mathcal{T} , given that H' displays these trees. It follows directly from the definition of the deletion AAF of a CNET that H and \mathcal{H} have the same deletion AAF.

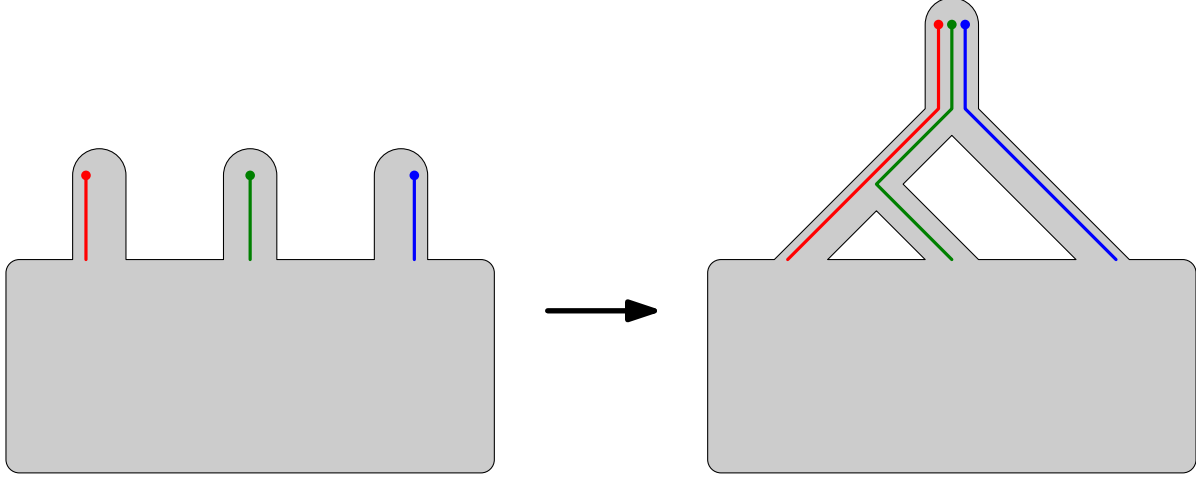


Figure 4: The third step of transforming a CNET into a hybridization network: combining multiple roots.

Now assume we are given a hybridization network H with hybridization number k for \mathcal{T} . Since H displays all trees in \mathcal{T} , we can choose a tree image $H(T)$, for every tree $T \in \mathcal{T}$, that includes the root of H . Then we set $\mathcal{H} = (H, \mathcal{E})$. \mathcal{H} satisfies conditions (i)–(v) of a CNET but may violate the remaining three conditions. Next we describe transformations that we apply to \mathcal{H} to ensure it satisfies these remaining conditions without introducing any violations of the conditions \mathcal{H} already satisfies and without increasing the hybridization number of \mathcal{H} . Thus, after applying these transformations, we obtain a CNET with hybridization number at most k for \mathcal{T} .

Condition (vi). Deleting all edges of H that are not contained in $\bigcup_{T \in \mathcal{T}} H(T)$ does not violate conditions (i)–(v) and establishes condition (vi). Since it also does not increase the hybridization number of \mathcal{H} , we obtain a network with hybridization number at most k that satisfies conditions (i)–(vi).

Condition (vii). As long as there is a *non-splitting* non-root node x , that is, a non-root node with only one child, we contract the edge e between x and its child in H and in every tree image $H(T) \in \mathcal{E}$ that includes e , and merge any parallel edges this may create. Each such contraction reduces the number of non-splitting non-root nodes by one, does not introduce any violations of conditions (i)–(vi), and does not increase the hybridization number of \mathcal{H} . Thus, we eventually obtain a network with hybridization number at most k that satisfies conditions (i)–(vii).

Condition (viii). By condition (vii), every non-root node of H is a split node. We call it a *true split node* if it also satisfies condition (viii), and a *fake split node* otherwise. We also call a true split node a *T-split node* if the tree image $H(T)$ contains both its child edges. The *weight* of a node x is the number of trees $T \in \mathcal{T}$ such that x is a *T-split node*. The *weight* of a path is the sum of the weights of the nodes on the path. Now we define the *potential* ϕ_x of a fake split node to be one plus the maximum weight of a path from a root to x . All other nodes have potential 0. The potential of the network is $\Phi := \sum_x \phi_x$, where the sum is taken over all nodes of the network. Since every fake split node has a positive potential,

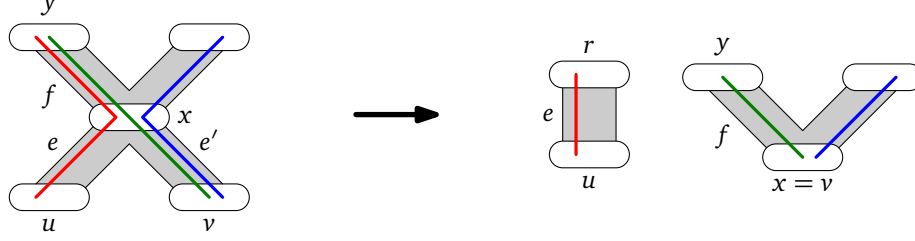


Figure 5: Eliminating a fake split node x below a root y .

a network has potential $\Phi = 0$ if and only if it contains no fake split node, that is, if and only if it satisfies condition (viii). Next we describe a transformation that decreases the potential of the network without increasing its hybridization number or introducing any violations of conditions (i)–(vii). Thus, after repeating this transformation as often as possible, we obtain a network with hybridization number at most k which satisfies conditions (i)–(viii), so it is a CNET with hybridization number at most k for \mathcal{T} .

While the network contains fake split nodes, there exists such a node x all of whose parents are true split nodes or roots. (Simply remove all roots, true split nodes, and leaves from H and choose x to be an in-degree-0 node of the resulting subgraph of H .) Since the colour sets of x 's child edges are disjoint and x has two child edges, one of these edges, e , must have exactly one colour: $C(e) = \{T\}$, for some $T \in \mathcal{T}$.² Let f be x 's parent edge that has colour T . By the choice of x , the top endpoint y of f is a root or a true split node.

If y is a root (see Figure 5), we remove T from the colour set of f , create a new root node r , change e 's top endpoint to r , remove f and its top endpoint y if the colour set of f is now empty and restore Condition (vii). It is easily verified that this does not introduce any violations of conditions (i)–(vi) and does not increase the hybridization number of the network. It also does not increase the potential of any node, and reduces the number of fake split nodes by one. The potential of the network therefore decreases.

If y is a true split node, we distinguish two cases. The first case is that y is a T' -split node, for some $T' \neq T$. This case is illustrated in Figure 6. Figure 6(a) depicts the subcase that y is not also a T -split node while Figure 6(b) depicts the subcase that y is also a T -split node. Both cases can be handled in a similar way.

Let g be the parent edge of y whose colour set includes T , and let f' be f 's sibling edge. We divide g into two edges g_t and g_b , with g_t above g_b , and denote their common endpoint by z . We remove T from the colour set of f , set $C(g_t) := C(g)$ and $C(g_b) := C(g) \setminus \{T\}$ if $T \notin C(f')$ and $C(g_b) := C(g)$ if $T \in C(f')$, change the top endpoint of e to z , remove all edges whose colour sets are now empty (this can only be f and g_b), and finally restore Condition (vii). Again, it is easily verified that this does not introduce any violations of conditions (i)–(vi) and does not increase the hybridization number of the network. It also does not increase the potential of any node and eliminates x from the network (because x has only one child edge e' after changing e 's top endpoint and hence e' is being contracted). The only

²This is the only place in the entire paper where we use that $|\mathcal{T}| = 3$. All other arguments are easily seen to generalize to more than 3 trees. Alas, this argument is crucial because our algorithm does not work without condition (viii).

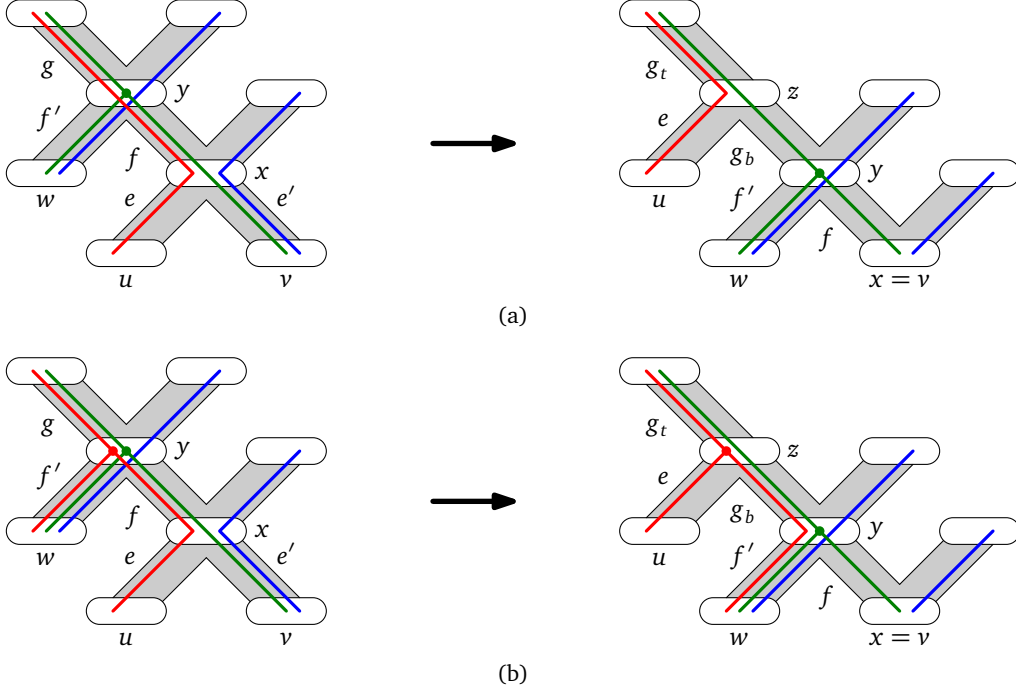


Figure 6: Eliminating a fake split node x below a split node y that is a T' -split node with T' not equal to the colour T (red) of the monochromatic edge below x . (a) y is not a T -split node. (b) y is a T -split node.

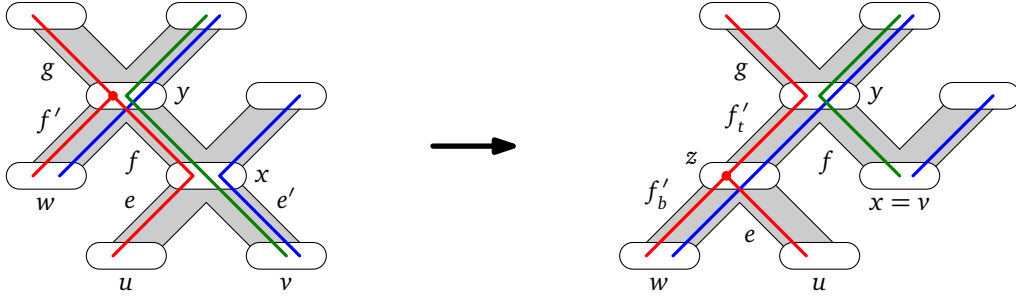


Figure 7: Eliminating a fake split node x below a split node y that is not a T' -split node for any T' that is not equal to the colour T (red) of the monochromatic edge below x .

new node is z . If $T \in C(f')$, then z is a true split node and its contribution to the network's potential is 0. Thus, since x is eliminated from the network, the network's potential decreases. If $T \notin C(f')$, then z is a fake split node. However, its potential ϕ_z is less than ϕ_x because for each a path from the root to z in the modified network there is a corresponding path from the root to x in the original network that has bigger weight because it contains true split node y . Thus, once again, the potential of the network decreases.

Finally, if y is not a T' -split node for any $T' \neq T$ (see Figure 7), it must be a T -split node. Let, as before, g be the parent edge of y whose colour set includes T , and let f' be f 's sibling edge. We subdivide f' into f'_t and f'_b where f'_t is above f'_b and let z be the newly created node. Change the top endpoint of e to z , remove T from the colour set of f , remove all edges whose colour sets are now empty

and restore Condition (vii). This transformation maintains conditions (i)–(vi) and does not increase the hybridization number of the network. It eliminates fake split node x from the network (because it has only one child edge e' after changing the top endpoint of e) and makes y a fake split node. However, the potential of y in the modified network is less than the potential of x in the original network. To see this, let P be a path of maximum weight from a root to y . Then the potential of y in the modified network is one plus the weight of P . In the original network, the path P extended by the edge (y, x) is then a path from a root to x , and its weight is one higher than the weight of P because it contains the true split node y . Hence, the potential of x in the original network is at least one higher than the potential of y in the modified network. Since the potential of all other nodes remains the same or decreases, the potential of the network decreases.

Let $\mathcal{H} = (H', \mathcal{E}')$ be the CNET eventually obtained by the above transformations and let H be the original hybridization network. It remains to show that \mathcal{H} and H have the same deletion AAF. Observe that any node and edge of any component of the deletion AAF of H is contained in any embedding of each of the input trees. Using this observation, it can easily be checked that none of the modifications above alters the deletion AAF. In particular, the only case in the modifications for satisfying Condition (viii) that involves an edge that is used by all three embeddings is the case depicted in Figure 6(b). Also the modification for this case does not alter the deletion AAF because the AAF component containing node w in the figure is not altered by the modification. \square

By Lemma 3, it is sufficient if our algorithm can construct a CNET of the three input trees. In the next section, we show how to do this.

5 Reconstructing a canonical network

Let $\mathcal{H} = (H, \mathcal{E})$ be a CNET for \mathcal{T} , let F be its deletion AAF, and let k be its hybridization number. For each tree $T \in \mathcal{T}$, let $I(T)$ be the set of nodes in T that do not belong to any path between two leaves x and y in the same AAF component (considering the root as a leaf). In a sense, these nodes are “invisible” in F . The *extended AAF* F^* of \mathcal{T} is defined as $F \cup I$, where $I := \bigcup_{T \in \mathcal{T}} I(T)$. We will refer to the elements of F^* as *components*. Let C be a component of F^* . Hence, C is either an AAF component or a vertex in I . If C is an AAF component, then r_C denotes the root of C . If C is a vertex of I , then r_C is equal to vertex C . In either case, we will refer to r_C as the *root* of component C .

By the following lemma, the size of I is at most $3(k - 1)$, when \mathcal{T} contains at most three trees.

Lemma 4. *For any $T \in \mathcal{T}$, $|I(T)| \leq k - 1$.*

Proof. Let $T \in \mathcal{T}$. Since the root of T is a leaf after omitting directions, we can see T as an unrooted tree. Since F is an AAF of \mathcal{T} and $T \in \mathcal{T}$, we know that F can be obtained by deleting a set E^* of k edges from T and then taking the partition of the leaves induced by the resulting connected components. Let \mathcal{C} be the partition of the vertices of $I(T)$ induced by the connected components of the subgraph of T induced by $I(T)$. For each $C \in \mathcal{C}$, let $\delta(C)$ denote the set of edges with exactly one endpoint in C . Then, by the definition of $I(T)$, at most one of the edges in $\delta(C)$ is not contained in E^* . Hence, from

the $|C| + 2$ edges in $\delta(C)$, at least $|C| + 1$ edges are in E^* . Moreover, since T is a tree, at most $|\mathcal{C}| - 1$ edges can be in $\delta(C) \cap \delta(C')$ for two different $C, C' \in \mathcal{C}$. Hence,

$$|E^*| \geq \sum_{C \in \mathcal{C}} (|C| + 1) - (|\mathcal{C}| - 1) = |I(T)| + 1.$$

Since $|E^*| = k$, the lemma follows. \square

We will construct \mathcal{H} from F^* and \mathcal{T} with the help of a *guess* of the structure of \mathcal{H} . In particular we construct \mathcal{H} by “gluing together” the components of F^* . Our guess concerns how this gluing is to be done. The root r_C of every component $C \in F^*$ has a unique image $H(r_C)$ in H : If $C \in F$, this is true because F is the deletion AAF of \mathcal{H} . If $r_C = C \in I(T)$, for some $T \in \mathcal{T}$, r_C is a split node of T and, thus, has a unique image $H(r_C)$ in $H(T)$ which is also a node of H . Our guess for r_C defines the “wiring” of the in-edges of $H(r_C)$; we call it the *wiring guess* for r_C . Since the colour sets of the parent edges of every node in H are disjoint, $H(r_C)$ has between one and three parent edges. The first part of the wiring guess for r_C is the number of parent edges of $H(r_C)$. The second part of the wiring guess for r_C is which of these in-edges is included in which tree image. Finally, observe that the top endpoint x of each parent edge of $H(r_C)$ must once again be a T' -split node, for at least one $T' \in \mathcal{T}$. The third part of the wiring guess for r_C determines such a tree T' for each parent edge of $H(r_C)$. We will assume without loss of generality that any two nodes $r_C, r_{C'}$ for which $H(r_C)$ and $H(r_{C'})$ have a common parent x both guess the same tree T' as the tree for which x is a T' -split node.

First consider a component root r_C that is a vertex in $I(T)$ for some $T \in \mathcal{T}$. Note that (i) at least one parent edge of $H(r_C)$ must have colour T because $H(r_C)$ is a non-root node of $H(T)$, and (ii) the top endpoint of a parent edge e of $H(r_C)$ can be a T' -split node only for trees T' such that $T' \in C(e)$. This gives the 17 possible wiring guesses for $r(C)$ shown in Figure 8.

If r_C is an AAF root, the set of possible wiring guesses is more restricted. Since $H(r_C)$ is contained in every tree image, the only valid wiring guesses are the ones where the union of the colour sets of the parent edges of $H(r_C)$ contains all three colours. This reduces the number of possible wiring guesses for AAF roots to 10 (see again Figure 8). Finally note that, when r_C is the root ρ of the trees, there is only a single wiring guess.

Our *guess* \mathcal{G} of the structure of \mathcal{H} consists of the wiring guesses for all roots r_C , $C \in F^*$.

Since we have 17 wiring guesses to choose from for each component in I , and 10 wiring guesses to choose from for each component in F that does not contain the tree root ρ , there are $10^{|F|-1} \cdot 17^{|I|}$ possible guesses \mathcal{G} . Since $|F| \leq k + 1$ and $|I| \leq 3(k - 1)$ by Lemma 4, the number of possible guesses \mathcal{G} is bounded by $10^k \cdot 17^{3(k-1)} = 49130^k / 4913$.

Our algorithm considers each guess in \mathcal{G} in turn and attempts to construct a CNET \mathcal{H} from $(\mathcal{G}, F^*, \mathcal{T})$ in polynomial time. We call $(\mathcal{G}, F^*, \mathcal{T})$ the CNET’s *description*. To establish our algorithm’s correctness, we prove that, given the description $(\mathcal{G}, F^*, \mathcal{T})$ of a CNET \mathcal{H} for \mathcal{T} , our algorithm succeeds in constructing a CNET \mathcal{H}' for \mathcal{T} with description $(\mathcal{G}, F^*, \mathcal{T})$ that differs from \mathcal{H} only in insignificant details and in particular has the same hybridization number as \mathcal{H} . Our proof is divided into two lemmas. The first one shows that two CNETs with the same description are essentially the same. We make this

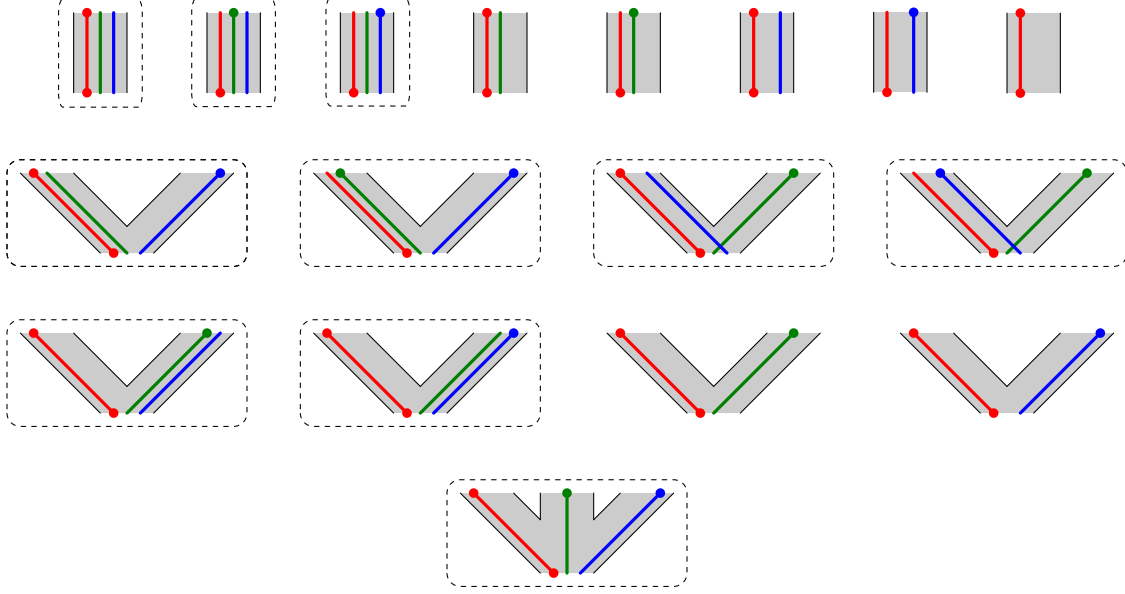


Figure 8: The 17 possible wiring guesses for the root r_C of a component C of the extended AAF F^* that is a node in $I(T)$, with T the red tree. The guess of the split node at the top of each edge is indicated by a dot. Valid guesses for a root r_C of an AAF component are indicated by dashed boxes around them.

precise below. The second one shows that, given the description of a CNET, we can construct a CNET with this description in polynomial time. Note that not every description $(\mathcal{G}, F^*, \mathcal{T})$ is necessarily a valid description of any CNET. If there is no CNET with the given description, then our algorithm will determine this polynomial time. The description of this algorithm is given in the proofs of Lemmas 5 and 6 below. Appendix A provides an example of the operation of the algorithm.

Given a CNET $\mathcal{H} = (H, \mathcal{E})$ for \mathcal{T} , we obtain a DAG \tilde{H} by contracting the image of every component of the deletion AAF F of H into a single node, keeping parallel edges this creates. We call \tilde{H} the *signature* of H . For an example, see Figure 12 in the appendix. It is easy to see that the node set of \tilde{H} is $\{H(r_C) \mid C \in F^*\}$ and that \tilde{H} has the same hybridization number as H . Note that two nodes $C \in I(T)$, $C' \in I(T')$ might map to the same node $H(r_C) = H(r'_C)$ of \tilde{H} . We label every node x of \tilde{H} with the set of roots r_C , $C \in F^*$, such that $x = H(r_C)$. We call roots that label the same node of \tilde{H} *buddies* (of each other). Note that roots of components of F have no buddies. We call two CNETs $\mathcal{H} = (H, \mathcal{E})$ and $\mathcal{H}' = (H', \mathcal{E}')$ *equivalent* if $\tilde{H} = \tilde{H}'$. In particular, two equivalent CNETs have the same hybridization number.

Before proving that CNETs with the same description are always equivalent, we introduce two useful notions which we will use in the two upcoming proofs.

We will use the following notion of the “attached subtrees” of an AAF component C in a tree T . For each edge (u, v) of T for which u is contained in $T(C)$ but v is not contained in $T(C)$, we say that the subtree of T induced by u, v and all nodes reachable from v is a subtree of T *attached* to C .

In addition, we will use the following notion of the “pendant subtrees” represented by an edge $e = (r_C, r_{C'})$ of the signature \tilde{H} . For each $T \in C(e)$, the subtree of T induced by $r_{C'}$, the parent of $r_{C'}$ and all

vertices reachable from $r_{C'}$ is said to be a *pendant subtree* of T represented by the edge e of \tilde{H} .

Lemma 5. *If two CNETs have the same description, then they are equivalent.*³

Proof. Given the description $(\mathcal{G}, F^*, \mathcal{T})$ of a CNET $\mathcal{H} = (H, \mathcal{E})$, we show how to uniquely reconstruct \tilde{H} from this description. We define a DAG D whose nodes are the roots of components of F^* . There is an edge from a root r_C to a root $r_{C'}$ if and only if there exists a tree $T \in \mathcal{T}$ such that $r_{C'}$ is an ancestor of r_C in T and there exists no component root $r_{C''}$ that is an ancestor of r_C and a descendant of $r_{C'}$ in T . If this is the case, we call r_C a *direct descendant* of $r_{C'}$ in T . We incrementally construct a topological ordering of D , define U_i to be the first i nodes in this topological ordering, D_i to be the subgraph of D induced by U_i and $V_i := \{H(r_C) \mid r_C \in U_i\}$. \tilde{D}_i is the subgraph of D induced by all nodes of D not in U_i . Let the partial signature \tilde{H}_i be the graph obtained from the subgraph of \tilde{H} induced by V_i by, for each parent edge e in \tilde{H} of a node $H(r_C) \in V_i$ whose top endpoint does not belong to V_i , adding a new root with an edge \tilde{e} from this new root to $H(r_C)$ and giving \tilde{e} the same colour set as e (which is determined by the set of guesses). For an example, see Figure 12 in the appendix. We will call an edge whose top endpoint is a root a *root edge*.

We assign these new roots colours, where the colour of the top endpoint of edge \tilde{e} is T if the bottom endpoint of \tilde{e} guesses that the top endpoint of e is a T -split node.

We will show that, from D_i, \tilde{H}_i and \mathcal{G} , we can determine a unique nonempty set of nodes U' of \tilde{D}_i , the set U^+ of all buddies of nodes in U' and an extended partial signature \tilde{H}_j corresponding to the graph D_j induced by $U_j := U_i \cup U' \cup U^+$.

Once $D_i = D$, we thus obtain $\tilde{H}_i = \tilde{H}$, that is, we are able to reconstruct the signature \tilde{H} only given the description of \mathcal{H} .

We initialize the reconstruction by defining D_0 and \tilde{H}_0 to be empty digraphs. Now consider an iteration with input digraph D_i and partial signature \tilde{H}_i . For an in-degree-0 node r_C of \tilde{D}_i that belongs to $I(T)$, for some $T \in \mathcal{T}$, observe that both child edges of $H(r_C)$ in \tilde{H} have corresponding root edges in \tilde{H}_i (by the construction of \tilde{H}_i). For such a node r_C , we call r_C *free* if the top endpoints of both of these root edges are coloured T . For an in-degree-0 node r_C of \tilde{D}_i that is a root of a component of F , we say that r_C is *free* if, for every in-edge e of r_C in D , the corresponding root edge \tilde{e} of \tilde{H}_i has the property that, for every $T \in C(\tilde{e})$, the pendant subtree of T represented by \tilde{e} is attached to the AAF component C in T . All other nodes of \tilde{D}_i are non-free. We claim that at least one of the nodes in \tilde{D}_i is free and that every free node can determine its buddies in \tilde{H} and can augment the partial signature \tilde{H}_i to \tilde{H}_j .

Consider a node x of the signature \tilde{H} that does not belong to \tilde{H}_i and all of whose children do belong to \tilde{H}_i . Since \tilde{H} is a DAG, such a node exists. Node x is a T -split node, for some $T \in \mathcal{T}$, and is thus the image $H(r_C)$ of the root r_C of a component in $F \cup I(T)$.

First consider the case that $C \in F$. Observe that, because \tilde{H} is the signature of a CNET \mathcal{H} for \mathcal{T} , the subtrees attached to C in all input trees are exactly the pendant subtrees represented by the child edges of $H(r_C)$ in \tilde{H} . Since these child edges are root edges of \tilde{H}_i , r_C is free. Node r_C can trivially identify its set of buddies because it has no buddies besides itself. We obtain D_j by adding r_C to D_i and obtain \tilde{H}_j

³The reverse statement does not hold.

from \tilde{H}_i by locating the root edges of \tilde{H}_i that correspond to child edges of $H(r_C)$ in \tilde{H} , merging the top endpoints of these root edges to a single node $H(r_C)$, and adding new root edges entering $H(r_C)$ according to r_C 's wiring guess. This gives a new graph $D_j \supset D_i$ and the corresponding graph \tilde{H}_j .

Now consider the case that $r_C \in I(T)$. In this case, both child edges of r_C in \tilde{H} correspond to root edges of \tilde{H}_i . Since these two edges have a common top endpoint in \tilde{H} , their top endpoints in \tilde{H}_i must have the same colour T' . This implies that $H(r_C)$ is a T' -split node and is thus the image $H(r_{C'})$ of a node $r_{C'} \in I(T')$. Since the child edges of $H(r_{C'})$ in \tilde{H} belong to \tilde{H}_i , both in-neighbours of $r_{C'}$ in D belong to D_i . Thus, $r_{C'}$ is free. Since $H(r_C) = H(r_{C'})$, we may assume that $r_{C'} = r_C$ (because in at least one set of guesses this is the case). We obtain \tilde{H}_j from \tilde{H}_i by locating the two root edges of \tilde{H}_i that correspond to child edges of $H(r_C)$ in \tilde{H} , merging the top endpoints of these edges to a single node $H(r_C)$, and adding new root edges entering $H(r_C)$ according to r_C 's wiring guess. Now consider the child edges of $H(r_C)$ in \tilde{H}_j . For each tree $T' \in \mathcal{T}$ such that these edges are both coloured T' , $H(r_C)$ is the image $H(r_{C'})$ of a node in $I(T')$. The nodes of \bar{D}_i that satisfy this condition are the buddies of r_C , and we add them to D_i along with r_C to obtain D_j .

We have shown that every node $H(r_C)$ of \tilde{H} whose children belong to \tilde{H}_i corresponds to a free node r_C in \bar{D}_i and that r_C can determine its set of buddies and can construct the extended partial signature \tilde{H}_j corresponding to the digraph $D_j \supset D_i$ obtained by adding these buddies to D_i . If r_C is an AAF root, it has no buddies besides itself. If $r_C \in I(T)$, for some $T \in \mathcal{T}$, observe that none of its buddies is free because it belongs to a tree $T' \neq T$ but the top endpoints of the child edges of $H(r_C)$ have colour T in \tilde{H}_i . Thus, to prove that every free node can determine its set of buddies and can construct \tilde{H}_j from \tilde{H}_i , it suffices to show that there is no free node r_C such that $H(r_C)$ has an in-neighbour not in \tilde{H}_i .

Assume there exists such a node r_C . Then r_C has in-degree 0 in \bar{D}_i because otherwise it is not free. First assume r_C is the root of a component in F . Let $H(r_{C'})$ be an in-neighbour of $H(r_C)$ in \tilde{H} that does not belong to \tilde{H}_i , and let T be an arbitrary colour T in the colour set of the edge e between $H(r_C)$ and $H(r_{C'})$ in \tilde{H} . Since r_C has in-degree 0 in \bar{D}_i , the pendant subtree of T represented by e is also represented by some root edge f of \tilde{H}_i . Thus, \tilde{H} contains a unique path from e to f and every node on this path is a T' -split node, for some $T' \neq T$. This implies in particular that the top endpoint of edge f is a T' -split node, $T, T' \in C(f)$, and the pendant subtree of T' represented by f is not a subtree attached to C in T' . Thus, since f represents the same in-edge of r_C in D as e , r_C is not free, a contradiction.

If $r_C \in I(T)$, for some tree $T \in \mathcal{T}$, we choose an in-neighbour $H(r_{C'})$ and a root edge f of \tilde{H}_i as in the case when $r_C \in F$. Since the top endpoint of edge f is not a T -split node, its colour in \tilde{H}_i must be $T' \neq T$. Thus, since f is the root edge of \tilde{H}_i representing one of the in-edges of r_C in D , r_C is not free, again a contradiction. \square

By Lemma 5, it suffices to provide a polynomial-time algorithm that decides whether there exists a CNET with description $(\mathcal{G}, F^*, \mathcal{T})$ and, if so, construct any such CNET. Our next lemma states that such an algorithm exists.

Lemma 6. *Given a description $(\mathcal{G}, F^*, \mathcal{T})$, it takes polynomial time to decide whether there exists a CNET with this description and, if so, to construct such a CNET.*

Proof. The proof of Lemma 5 already specifies an algorithm for constructing the signature \tilde{H} of a CNET $\mathcal{H} = (H, \mathcal{E})$ with description $(\mathcal{G}, F^*, \mathcal{T})$, if such a CNET exists, and this algorithm is clearly implementable in polynomial time. This algorithm provides a first test that can be used to reject invalid descriptions: If the algorithm reaches an iteration where \tilde{D}_i is non-empty but none of its nodes is free, then the description is invalid because the proof of Lemma 5 shows that, if $(\mathcal{G}, F^*, \mathcal{T})$ is the description of a CNET, then there exists a free node in each iteration. Thus, the algorithm aborts and rejects the description. If the algorithm does not reject the description, its output is a signature \tilde{H} that respects all wiring guesses in \mathcal{G} . However, this signature may not correspond to a network H that displays all input trees. Next we present a polynomial-time algorithm for constructing $\mathcal{H} = (H, \mathcal{E})$ from \tilde{H} , F^* , and \mathcal{T} or determining that no CNET $\mathcal{H} = (H, \mathcal{E})$ with description $(\mathcal{G}, F^*, \mathcal{T})$ exists.

The nodes of \tilde{H} are of two types: images of AAF roots and images of nodes in I . To obtain a CNET $\mathcal{H} = (H, \mathcal{E})$ with description $(\mathcal{G}, F^*, \mathcal{T})$ from \tilde{H} , we let H initially be equal to \tilde{H} and will replace each AAF root image with the AAF component it represents. We process these AAF root images bottom-up, that is, in reverse topological order.⁴ Consider such an image $H(r_C)$ of the root r_C of an AAF component C , and let E be the set of child edges of $H(r_C)$ in \tilde{H} . We remove the edges in E from \tilde{H} and attach C below r_C , setting $C(e) = \mathcal{T}$ for every edge e of C . Our goal now is to reattach the edges in E to edges of C so that, for all $T \in \mathcal{T}$, the descendant edges of $H(r_C)$ with colour T form an image of the pendant subtree of T with root r_C . For each tree T , observe that the edges in E coloured T represent the subtrees attached to C in T . We need to attach these edges to C in H so that each edge branches off the same edge of C as in T and edges that branch off the same edge of C in T do so in the same order as in T .

First we test every edge $e \in E$ whether e branches off the same edge of C in every tree $T \in C(e)$. If this is the case for all edges $e \in E$, then we can partition E into subsets E_f , one per edge f of C , such that all edges in E_f branch off edge f . If there is an edge e that branches off some edge f in a tree $T \in C(e)$ and off a different edge f' in a tree $T' \in C(e)$, it is impossible to attach this edge to C in a way that satisfies both constraints. Since the edges of T and T' represented by e are determined by \tilde{H} , which in turn is uniquely defined by $(\mathcal{G}, F^*, \mathcal{T})$, there is thus no network \mathcal{H} with description $(\mathcal{G}, F^*, \mathcal{T})$, so the algorithm aborts and reports that there is no such network.

Given the partition of E into subsets E_f such that the edges in E_f branch off edge f , we need to attach the edges in each such set E_f in an ordering consistent with the input trees. Let $E_{f,T}$ be the subset of edges in E_f that have colour T . Tree T determines the ordering in which these edges need to be attached to f . We define a DAG D_f whose nodes represent the edges in E_f , and which has an edge (g, g') precisely if there is a tree T such that $g, g' \in E_{f,T}$ and g branches off f in T above g' . There exists an ordering in which to attach the edges of E_f to f so that the ordering constraints imposed by all trees in \mathcal{T} are satisfied if and only if D_f is acyclic. Moreover, if D_f is indeed acyclic, then a topological ordering of D_f provides a valid ordering in which the edges can be attached. Thus, we can test whether such an ordering exists and, if so, compute such an ordering in time $O(|E_f|)$ per edge f . If no such ordering exists, the algorithm once again aborts and reports that there is no CNET with description $(\mathcal{G}, F^*, \mathcal{T})$. If we can find a correct ordering of the edges attached to each edge f of C , then the replacement of r_C

⁴This is not really essential, but it simplifies the description of the algorithm.

with C in this manner results in a network where all descendant edges of $H(r_C)$ with colour T form an image of the pendant subtree of T with root r_C , for all $T \in \mathcal{T}$. Thus, after replacing each AAF root r_C with its corresponding component C in this fashion, we obtain a CNET $\mathcal{H} = (H, \mathcal{E})$ with description $(\mathcal{G}, F^*, \mathcal{T})$. \square

6 Conclusions

It seemed natural enough to try to reconstruct a hybridization network from the corresponding AAF and to guess the wiring structure that determines how the AAF components need to be glued together to achieve this. In the case of two trees, no guessing is required: the AAF carries the necessary information. For more than two trees, there are only a constant number of choices for the wiring of the root of each component, so with $O(k)$ components, one obtains an $O(c^k \cdot \text{poly}(n))$ -time algorithm. Unfortunately, guessing the wiring structure of AAF components turned out not to be enough, even for three trees, because there are examples of three input trees such that every optimal network displaying these trees contains an *invisible component*: a group of nodes that are isolated from all taxa once all hybridization edges are deleted, see Figure 1 in the introduction. We call these components invisible because they are not represented in any form in the AAF.

Guessing the number and structure of these invisible components seems extremely challenging. In this paper, we showed that one can get away without having to guess these components in the case of three trees because, for three trees, these components are not invisible after all: They may not be represented in the AAF, but they are present as nodes in the three input trees, at least as long as we consider only canonical networks (and we have shown that we may do this without loss of generality). This is the key to our $O(c^k \cdot \text{poly}(n))$ -time algorithm for three trees. Unfortunately, it appears that we simply scraped by. While the framework of our algorithm extends to more than three trees, it seems that already for four trees, there are input instances where the optimal network includes truly invisible nodes: nodes whose only purpose is to change the way in which edges of the tree images are braided together along network edges, see Figure 9. Thus, the main open problem is to discover structural properties that, while unlikely to eliminate the need to guess the existence of these braiding structures in the network altogether, at least limit the number of possible guesses to be explored.

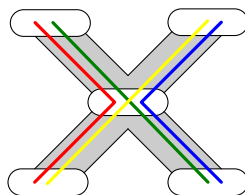


Figure 9: A “truly invisible” node of a network with four trees embedded in it. None of the four trees branches in this network node.

Another interesting question that arises from our work is whether guessing the wiring of the extended AAF components as part of the reconstruction is necessary at all, at least in the case of three trees. Since

all these components are visible in the input trees, it could be possible that one can construct the entire network directly from the AAF, that is, as in the case of two trees, the hard core of the problem is finding the right AAF. However, we conjecture that this is not the case: i.e., that even given the deletion AAF of an optimal hybridization network for the three input trees, it remains NP-hard to find the network.

References

- [1] Eric Bapteste, Leo van Iersel, Axel Janke, Scot Kelchner, Steven Kelk, James O. McInerney, David A. Morrison, Luay Nakhleh, Mike Steel, Leen Stougie, and James Whitfield. Networks: expanding evolutionary thinking. *Trends in Genetics*, 29(8):439 – 441, 2013.
- [2] M. Baroni, C. Semple, and M. Steel. A framework for representing reticulate evolution. *Annals of Combinatorics*, 8:391–408, 2004.
- [3] Mihaela Baroni, Stefan Grünewald, Vincent Moulton, and Charles Semple. Bounding the number of hybridisation events for a consistent evolutionary history. *Mathematical Biology*, 51:171–182, 2005.
- [4] Magnus Bordewich and Charles Semple. Computing the hybridization number of two phylogenetic trees is fixed-parameter tractable. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 4(3):458–466, 2007.
- [5] Magnus Bordewich and Charles Semple. Computing the minimum number of hybridization events for a consistent evolutionary history. *Discrete Applied Mathematics*, 155(8):914–928, 2007.
- [6] Zhi-Zhong Chen and Lusheng Wang. An ultrafast tool for minimum reticulate networks. *Journal of Computational Biology*, 20(1):38–41, 2013.
- [7] Rodney G. Downey and Michael R. Fellows. *Parameterized Complexity*. Springer-Verlag, 1999.
- [8] O. Gascuel and M. Steel, editors. *Reconstructing Evolution: New Mathematical and Computational Advances*. Oxford University Press, USA, 2007.
- [9] Daniel H. Huson, Regula Rupp, and Celine Scornavacca. *Phylogenetic Networks: Concepts, Algorithms and Applications*. Cambridge University Press, 2011.
- [10] S.M. Kelk, C. Scornavacca, and L.J.J. van Iersel. On the elusiveness of clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 9(2):517–534, 2012.
- [11] Steven Kelk and Celine Scornavacca. Towards the fixed parameter tractability of constructing minimal phylogenetic networks from arbitrary sets of nonbinary trees. arXiv:1207.7034 [q-bio.PE], 2012.
- [12] Steven Kelk and Celine Scornavacca. Constructing minimal phylogenetic networks from softwired clusters is fixed parameter tractable. *Algorithmica*, 68(4), 2014.

- [13] Steven Kelk, Leo van Iersel, Nela Lekić, Simone Linz, Celine Scornavacca, and Leen Stougie. Cycle killer... qu'est-ce que c'est? on the comparative approximability of hybridization number and directed feedback vertex set. *SIAM Journal on Discrete Mathematics*, 26(4):1635–1656, 2012.
- [14] Simone Linz and Charles Semple. Hybridization in non-binary trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 6(1):30–45, 2009.
- [15] David Morrison. *Introduction to phylogenetic networks*. RJR Productions, Uppsala, 2011.
- [16] R. Niedermeier. *Invitation to Fixed Parameter Algorithms (Oxford Lecture Series in Mathematics and Its Applications)*. Oxford University Press, USA, March 2006.
- [17] Teresa Piovesan and Steven Kelk. A simple fixed parameter tractable algorithm for computing the hybridization number of two (not necessarily binary) trees. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 10(1):18–25, 2013.
- [18] Charles Semple and Mike Steel. A supertree method for rooted trees. *Discrete Applied Mathematics*, 105(1-3):147–158, 2000.
- [19] Leo van Iersel and Steven Kelk. Kernelizations for the hybridization number problem on multiple nonbinary trees. 2014. arXiv:1311.4045 [cs.DM]. Preliminary version to appear in Workshop on Graph-Theoretic Concepts in Computer Science (WG 2014).
- [20] Leo van Iersel, Steven Kelk, Nela Lekić, and Leen Stougie. Approximation algorithms for nonbinary agreement forests. *SIAM Journal on Discrete Mathematics*, 28(1):49–66, 2014.
- [21] Leo van Iersel and Simone Linz. A quadratic kernel for computing the hybridization number of multiple trees. *Information Processing Letters*, 113(9):318 – 323, 2013.
- [22] Chris Whidden, Robert G. Beiko, and Norbert Zeh. Fixed-parameter algorithms for maximum agreement forests. *SIAM Journal on Computing*, 42(4):1431–1466, 2013.
- [23] Yufeng Wu. An algorithm for constructing parsimonious hybridization networks with multiple phylogenetic trees. *Journal of Computational Biology*, 20(10):792–804, 2013.
- [24] Yun Yu, R Matthew Barnett, and Luay Nakhleh. Parsimonious inference of hybridization in the presence of incomplete lineage sorting. *Systematic Biology*, 62(5):738–751, 2013.

A An Example of Constructing a Network from its Description

This appendix provides an example of the construction described in the proofs of Lemmas 5 and 6. Consider the description $(\mathcal{G}, F^*, \mathcal{T})$ in Figure 10. The CNET in Figure 13 has this description. We first show how to construct its signature using the construction in the proof of Lemma 5. This signature is shown in Figure 12. Then we discuss how to construct the CNET in Figure 13 from this signature. The hybridization network induced by this CNET is shown in Figure 14.

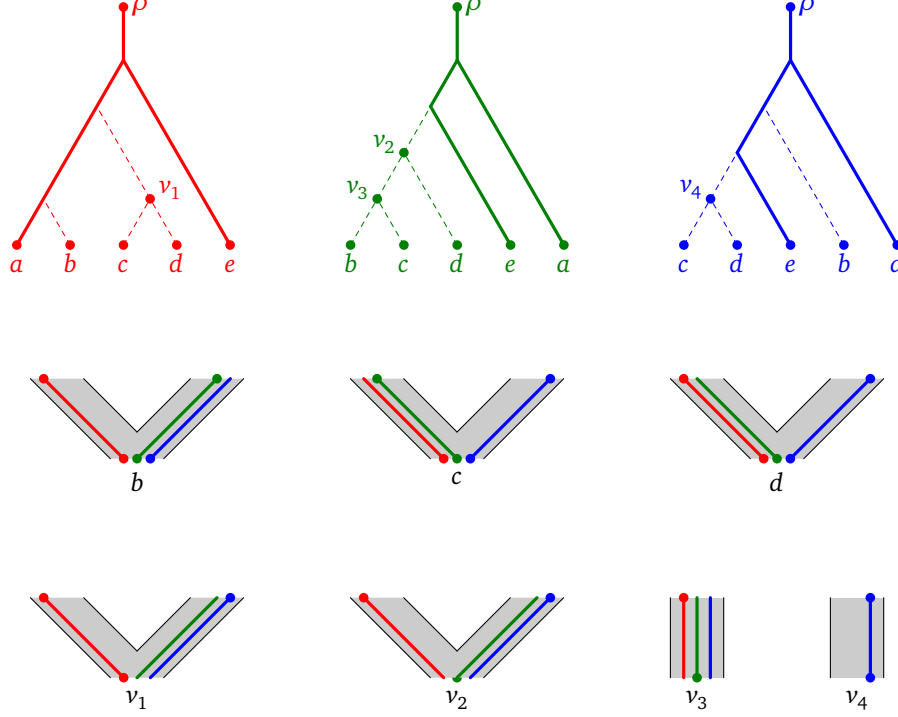


Figure 10: The input to the network reconstruction, including the AAF F (shown in bold in the three input trees), the set of “invisible nodes” $I = \{v_1, \dots, v_4\}$, and the wiring guesses for the resulting set of components of F^* . There is no guess for the component $\{a, e, \rho\}$ because it includes the root ρ of the three trees.

Lemma 5: Constructing the signature. The DAG D used in the construction of the signature \tilde{H} of any CNET with the description in Figure 10 is shown in Figure 11. This DAG represents the adjacency of components of F^* in the three input trees in \mathcal{T} .

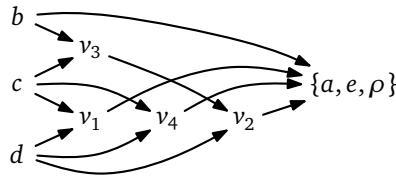


Figure 11: The DAG D used in the reconstruction of the signature \tilde{H} in Figure 12 from the description in Figure 10.

Before the first iteration, we have $\bar{D}_0 = D$. The only nodes with in-degree 0 in \bar{D}_0 are b , c , and d . They are all free because they have no in-edges at all. Assume we pick b as the first node to add to \tilde{H} . We create the node $H(b)$ and add parent edges according to b 's wiring guess to obtain the partial network \tilde{H}_1 in Figure 12. Since b is the root of an AAF component, it has no buddies, so D_1 has the node set $\{b\}$.

\bar{D}_1 has two nodes of in-degree 0, namely c and d . Both are again free. Assume we choose c as the next node to add to D_1 to obtain D_2 . Since c is again the root of an AAF component, it has no buddies,

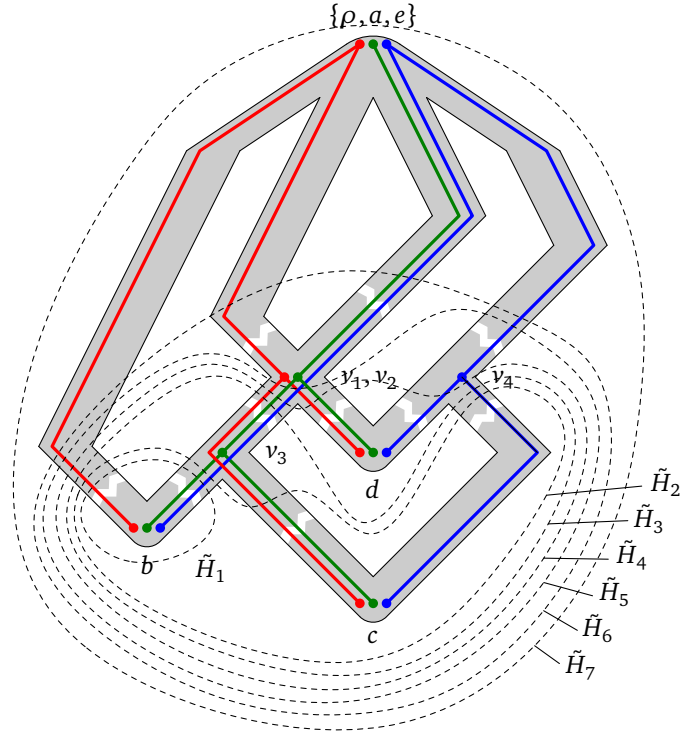


Figure 12: The signature \tilde{H} of any CNET with the description $(\mathcal{G}, F^*, \mathcal{T})$ shown in Figure 10. The partial signatures $\tilde{H}_1, \tilde{H}_2, \dots, \tilde{H}_7 = \tilde{H}$ constructed incrementally are indicated by dashed lines.

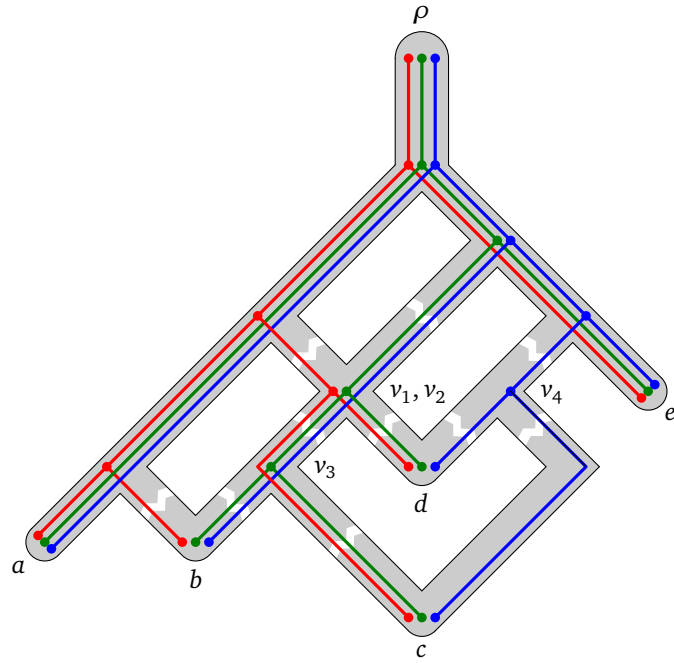


Figure 13: The CNET obtained from the signature in Figure 12 by expanding the components of F .

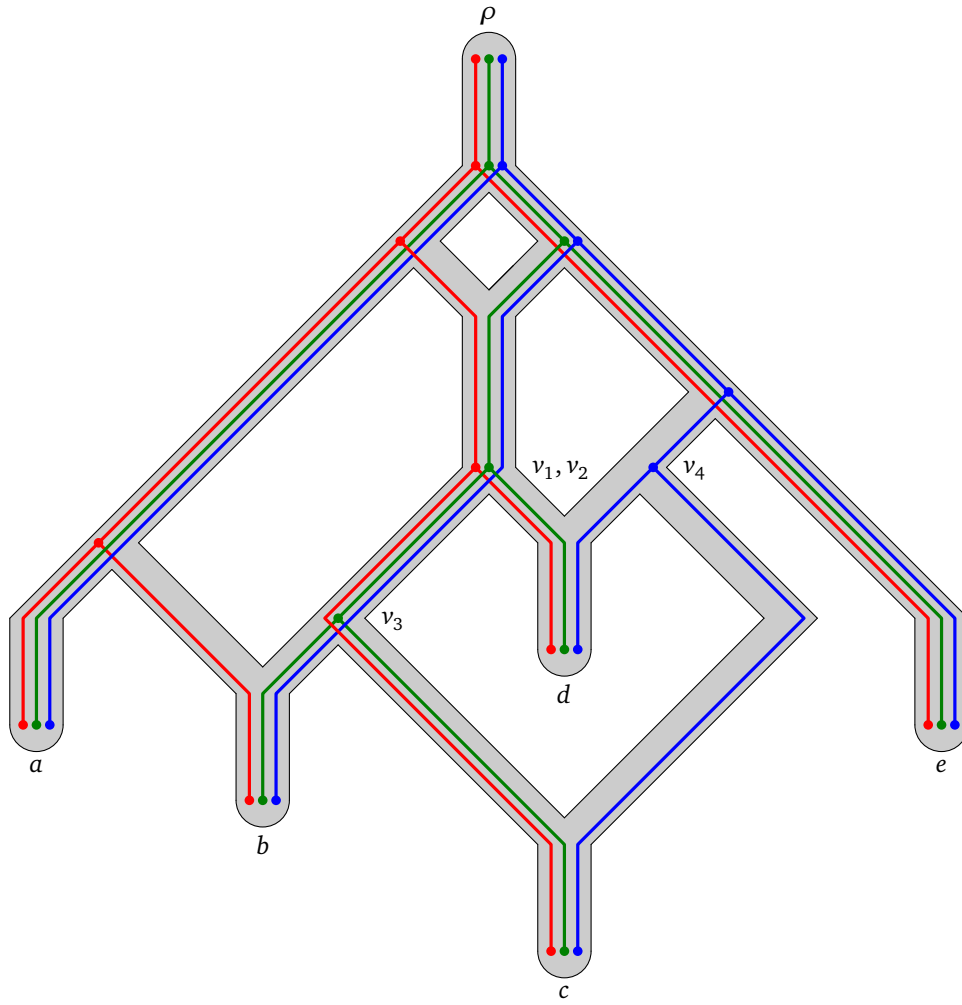


Figure 14: The hybridization network induced by the CNET in Figure 13, obtained by separating reticulations, split nodes, and leaves.

so the node set of D_2 is $\{b, c\}$. We construct the graph \tilde{H}_2 in Figure 13 from \tilde{H}_1 by creating a node $H(c)$ and adding parent edges of this node according to c 's wiring guess.

\tilde{D}_2 has two nodes of in-degree 0, namely d and v_3 . Both nodes are free: d is free because it is the root of an AAF component; v_3 is free because \tilde{H}_2 has two root edges above $H(b)$ and $H(c)$, which are children of v_3 in the green tree, and the top endpoints of both edges are coloured green by the wiring guesses for b and c . Let us assume we choose v_3 as the next node to add to D_2 to obtain D_3 . We create the node $H(v_3)$ by identifying the top endpoints of the two green parent edges of $H(b)$ and $H(c)$ and add a parent edge above $H(v_3)$ according to v_3 's wiring guess. This gives the graph \tilde{H}_3 shown in Figure 12. Since the only tree common to the colour sets of the parent edges of $H(b)$ and $H(c)$ is the green one, v_3 has no buddies. Thus, D_3 has node set $\{b, c, v_3\}$.

\tilde{D}_3 has d as its only in-degree-0 node, and d is free. We obtain D_4 by adding d to D_3 . Node d has no buddies because it is the root of an AAF component. To obtain \tilde{H}_4 from \tilde{H}_3 , we create the node $H(d)$ and add parent edges according to d 's wiring guess.

\tilde{D}_4 has three nodes of in-degree 0, namely v_1 , v_2 , and v_4 . The two child edges of v_1 are represented by red parent edges of $H(v_3)$ and $H(d)$ in \tilde{H}_4 . According to the wiring guesses for v_3 and d , their top endpoints are red. Since v_1 belongs to the red tree, v_1 is free. The same two edges also represent the child edges of v_2 . However, v_2 is green and, thus, is not free. Finally, the two child edges of v_4 are represented by the blue parent edges of $H(c)$ and $H(d)$, both of which have blue top endpoints according to the wiring guesses for c and d . Thus, v_4 is also free. Assume we choose v_4 as the next node to add to D_4 to obtain D_5 . We create the node $H(v_4)$ in \tilde{H}_5 by identifying the top endpoints of the blue parent edges of $H(c)$ and $H(d)$ and then create a parent edge of $H(v_4)$ according to v_4 's wiring guess. This produces the graph \tilde{H}_5 in Figure 12. Since the colour sets of the two child edges of $H(v_4)$ have only the blue tree in common, v_4 has no buddies and D_5 has node set $\{b, c, d, v_3, v_4\}$.

Nodes v_1 and v_2 are the only nodes of in-degree 0 in \tilde{D}_5 . Just as in \tilde{D}_4 , v_1 is free and v_2 is not. Thus, we choose v_1 as the node to add to D_5 to obtain D_6 . The two child edges of v_1 are represented by the red parent edges of $H(v_3)$ and $H(d)$ in \tilde{H}_5 . We identify their top endpoints to create the node $H(v_1)$ and add parent edges according to the wiring guess for v_1 . This produces the graph \tilde{H}_6 in Figure 12. Now observe that the two child edges of $H(v_1)$ in \tilde{H}_6 are also coloured green. Thus, we identify the node that is the parent of the two edges of the green tree represented by these child edges, which is node v_2 . Node v_2 becomes a buddy of v_1 and is added to D_5 along with v_1 to obtain D_6 . Thus, D_6 has node set $\{b, c, d, v_1, v_2, v_3, v_4\}$. Note that making v_1 and v_2 buddies does not create any conflicts because they both have the same wiring guess in \mathcal{G} .

Finally, the only node remaining in \tilde{D}_6 is $\{a, e, \rho\}$. The root edges of \tilde{H}_6 represent exactly the set of pendant edges of the AAF component $\{a, e, \rho\}$ in the three input trees, so $\{a, e, \rho\}$ is free in \tilde{D}_6 . We create a node $H(\{a, e, \rho\})$ in \tilde{H}_7 by identifying the top endpoints of all root edges of \tilde{H}_6 . Since there is no wiring guess for $\{a, e, \rho\}$ in \mathcal{G} , we do not add any parent edges to $H(\{a, e, \rho\})$, and $\tilde{H}_7 = \tilde{H}$ is the final signature.

It is easily verified that we would have obtained the exact same signature had we chosen different nodes to add to D_i in iterations where \tilde{D}_i contained more than one free node.

Lemma 6: Expanding AAF components. In our example, the only non-trivial AAF component to be expanded is the component $\{a, e, \rho\}$. This component has two non-root edges. Let f_a be the parent edge of a , and let f_e be the parent edge of e in this component. In \tilde{H} , the node $H(\{a, e, \rho\})$ has four child edges: a red parent edge e_1 of $H(b)$, a red parent edge e_2 of $H(v_1) = H(v_2)$, a green-blue parent edge e_3 of $H(v_1) = H(v_2)$, and a blue parent edge e_4 of $H(v_4)$. Edges e_1 and e_2 attach to f_a in the red tree and do not represent any edges in any other trees, so we add them to E_{f_a} . Edge e_3 attaches to f_e in the green and the blue tree, so there is no conflict and we add it to E_{f_e} . Edge e_4 attaches to f_e in the blue tree and does not represent any edge in any other tree, so we add it to E_{f_e} .

The DAG D_{f_a} has two nodes representing edges e_1 and e_2 with an edge from e_2 to e_1 because e_2 attaches to f_a above e_1 in the red tree. A topological ordering of D_{f_a} places these edges in the order $\langle e_2, e_1 \rangle$, and this is the order in which we attach e_2 and e_1 to f_a .

The DAG D_{f_e} has two nodes representing edges e_3 and e_4 with an edge from e_3 to e_4 because e_3 attaches to f_e above e_4 in the blue tree. The green tree does not impose any conflicting ordering constraints because only edge e_3 belongs to this tree. A topological ordering of D_{f_e} places e_3 and e_4 in the order $\langle e_3, e_4 \rangle$, and this is the order in which we attach e_3 and e_4 to f_e . The result is the CNET shown in Figure 13.

Finally, the hybridization network induced by this CNET is shown in Figure 14.